Intelligent Resource Optimization in Cloud Computing Using Artificial Intelligence

(English Executive Summary)

Ali Samali

ISBN: 978-622-88253-5-9
Publication Year: 2025 (1404)
Publisher: Ettela'at Negar Publishing

Original Language: Persian
Document Type: Academic Summary for International Review

Executive Summary

This book studies how Artificial Intelligence (AI) can improve resource management in cloud computing systems. The main goal is to show how AI can help cloud providers use computing resources more efficiently while keeping good performance and reducing costs.

Cloud computing systems work in dynamic environments. Workloads change over time, and user demand is often unpredictable. Cloud providers must manage CPU, memory, storage, and network bandwidth carefully. Traditional rule-based methods are not always enough for large and complex cloud systems.

The book starts with basic concepts of cloud computing. It explains service models such as IaaS, PaaS, and SaaS. It also discusses virtualization, containers, load balancing, and auto-scaling. These topics help the reader understand how cloud resources are managed in practice.

A central part of the book focuses on using AI methods for resource optimization. It explains supervised learning, unsupervised learning, and reinforcement learning in a simple and structured way. The book shows how these methods can support:

- Dynamic resource allocation
- Workload prediction
- Auto-scaling decisions
- Task scheduling
- Energy efficiency in data centers
- Cost optimization

The book also discusses time-series forecasting for demand prediction and reinforcement learning for adaptive scaling. Optimization techniques such as genetic algorithms and swarm-based methods are introduced to improve workload distribution.

Practical examples are explained in the context of major cloud platforms, including Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP). The book describes how AI-based optimization can reduce latency, improve throughput, lower energy use, and control operational costs.

In addition to technical solutions, the book addresses real-world challenges. These include data quality issues, scalability limits, training requirements, and the balance between system performance and reliability.

The final chapters explore future trends such as edge computing, multi-cloud management, green computing, and AI-supported infrastructure automation.

Overall, this book provides a clear and structured framework for students, researchers, cloud engineers, and IT professionals who want to apply AI methods to improve efficiency and sustainability in cloud computing systems.

Table of Contents